City University of New York (CUNY)
CUNY Academic Works

Open Educational Resources | City College of New York

2018

Lecture: Intro to Data Science - ML 2 - Week Ten

Grant Long
CUNY City College

NYC Tech-in-Residence Corps

How does access to this work benefit you? Let us know!

More information about this work at: https://academicworks.cuny.edu/cc_oers/188
Discover additional works at: https://academicworks.cuny.edu

This work is made publicly available by the City University of New York (CUNY).
Contact: AcademicWorks@cuny.edu

City College, Fall 2018

Intro to Data Science

Week 10: Performance Evaluation and Ensemble Models

November 12, 2018

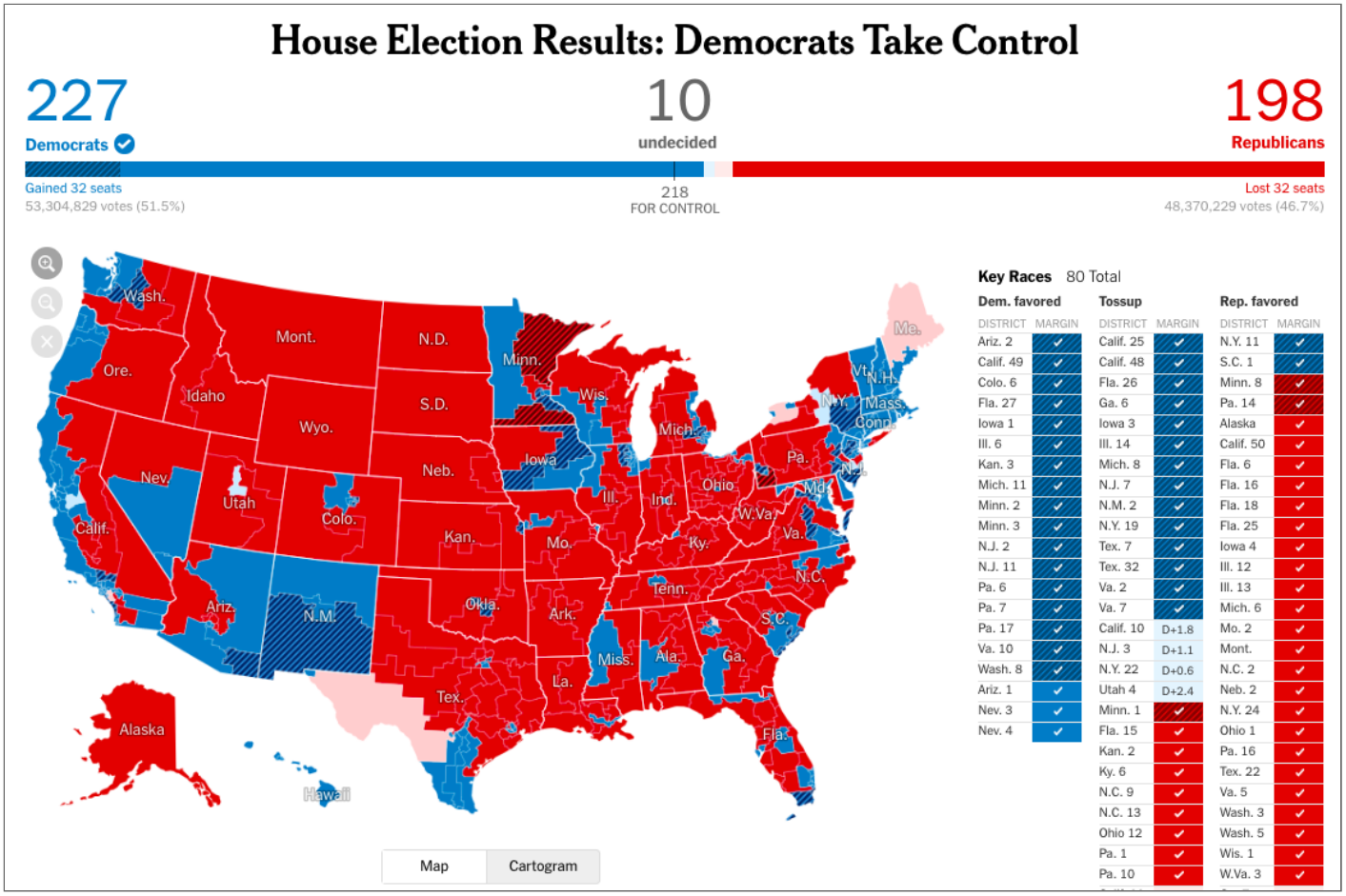
Today's Agenda

1. Project Notes
2. Model Fit
3. Ensemble Models

Part I

Evaluating Model Performance

Did we see this coming?

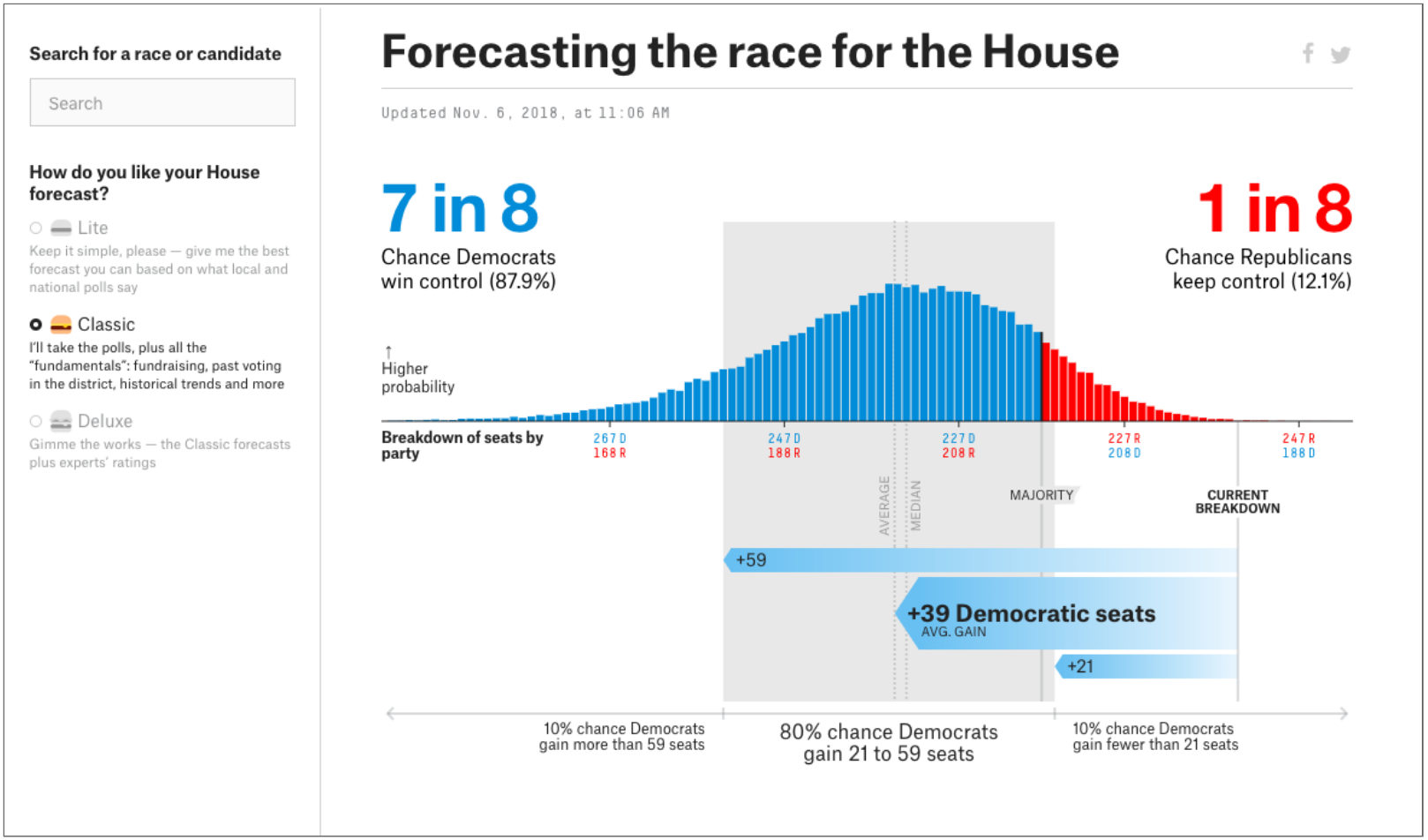


Democratic strategist David Axelrod predicted that poll performance "is going to prompt another round of soul-searching about whether and how you can poll accurately, because a lot of these races that were blowouts tonight or apparently blowouts tonight polled as tough races."

Democratic pollster Andrew Baumann called the pre-election polls "quite accurate, particularly for a midterm that ended up being totally different than any previous midterm."

Source.

Was this true?



538 Prediction Results

	Actual Dem	Actual Rep
Predicted Dem	215	9
Predicted Rep	16	195

Based on 538 deluxe model for 11/05/18.

What are the downsides of these measures?

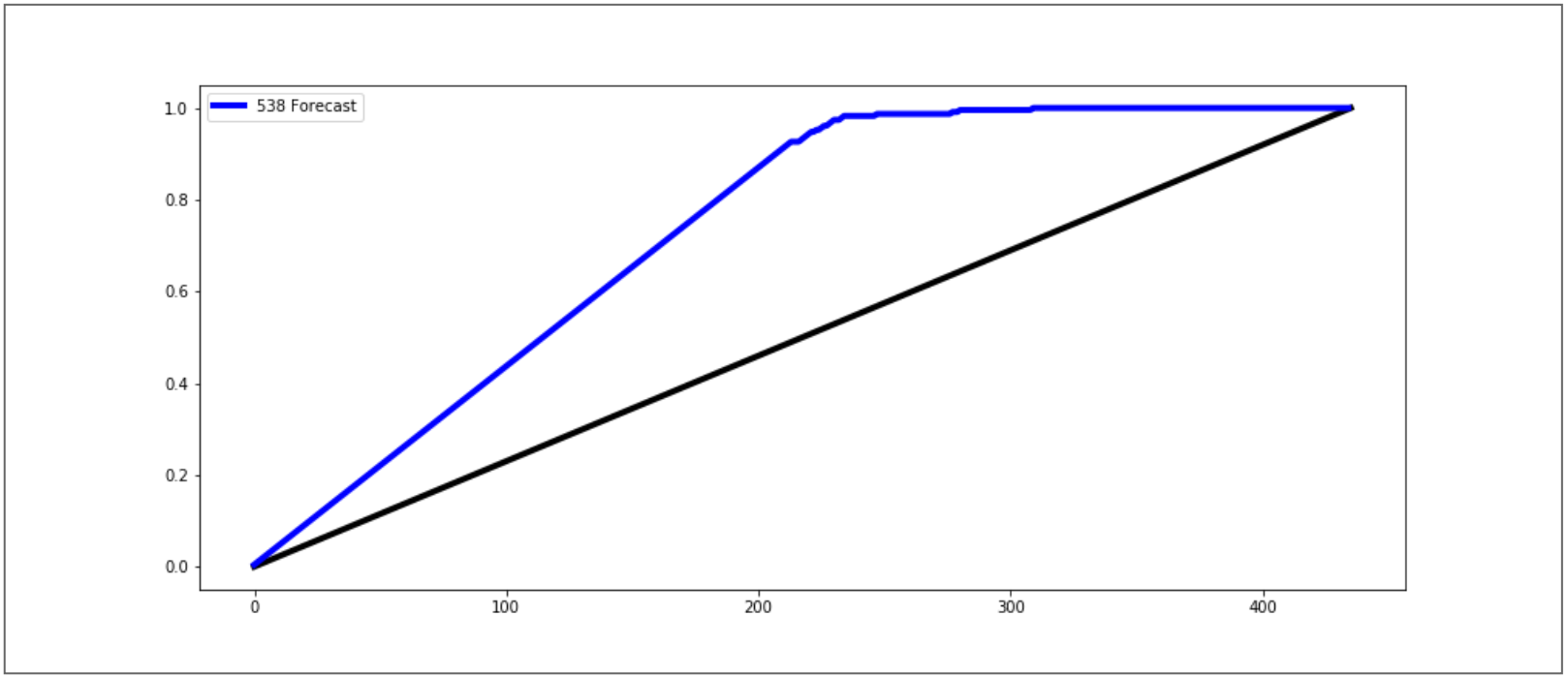
Lift Charts

Rank all observations by the predicted probability class, and chart the cummulative share of actual true values captured by the first x observations, where x ranges from 1 to the total number of observations.

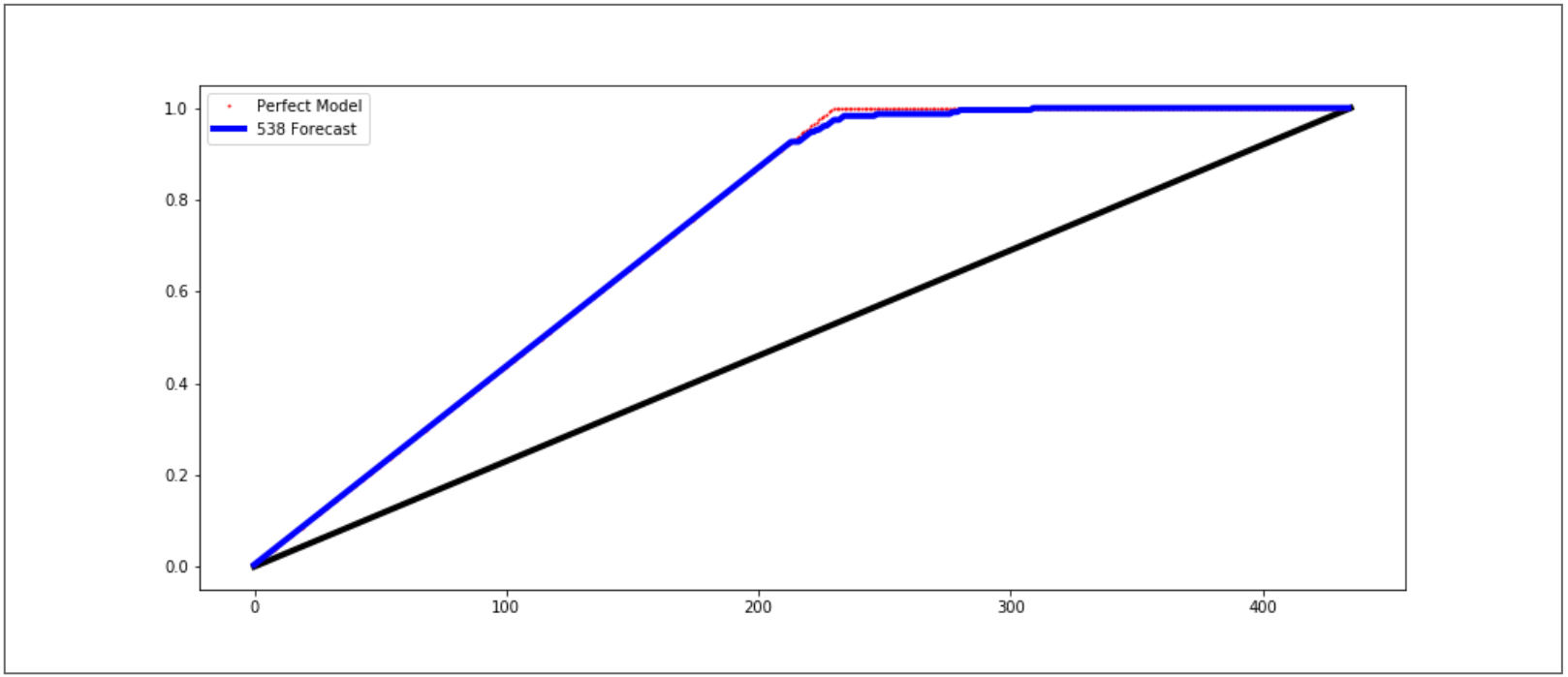
Demonstrates model's ability to outperform other (random) choices at positive prediction across decision thresholds.

For more details.

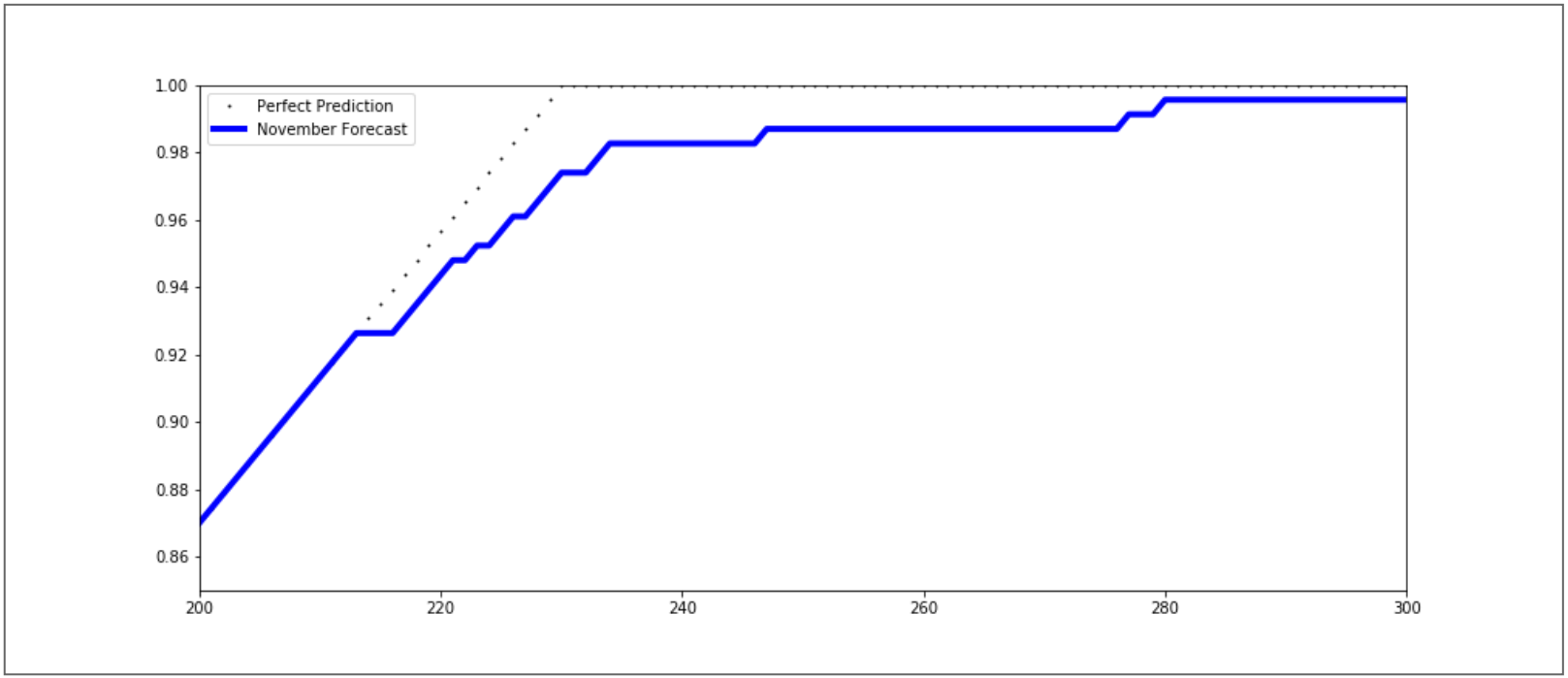
538 Lift



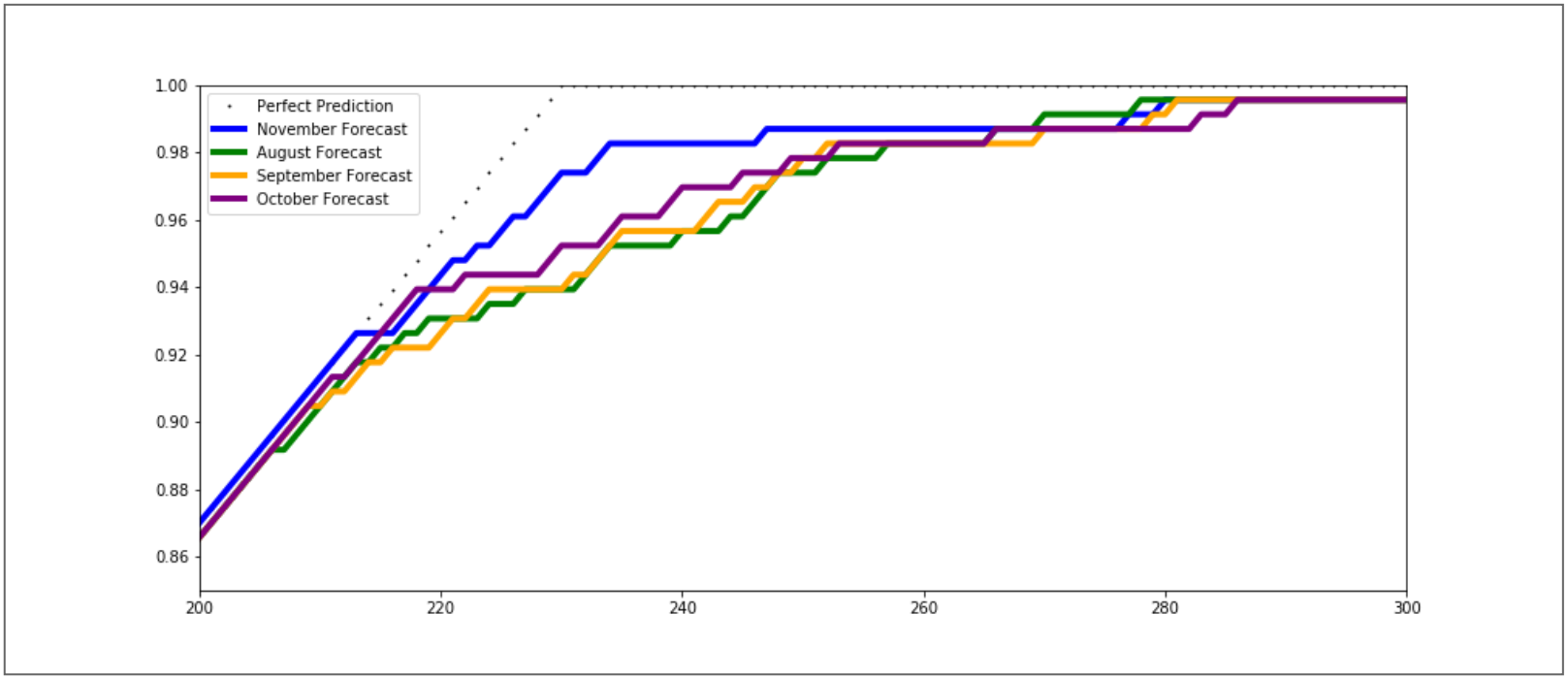
538 Lift



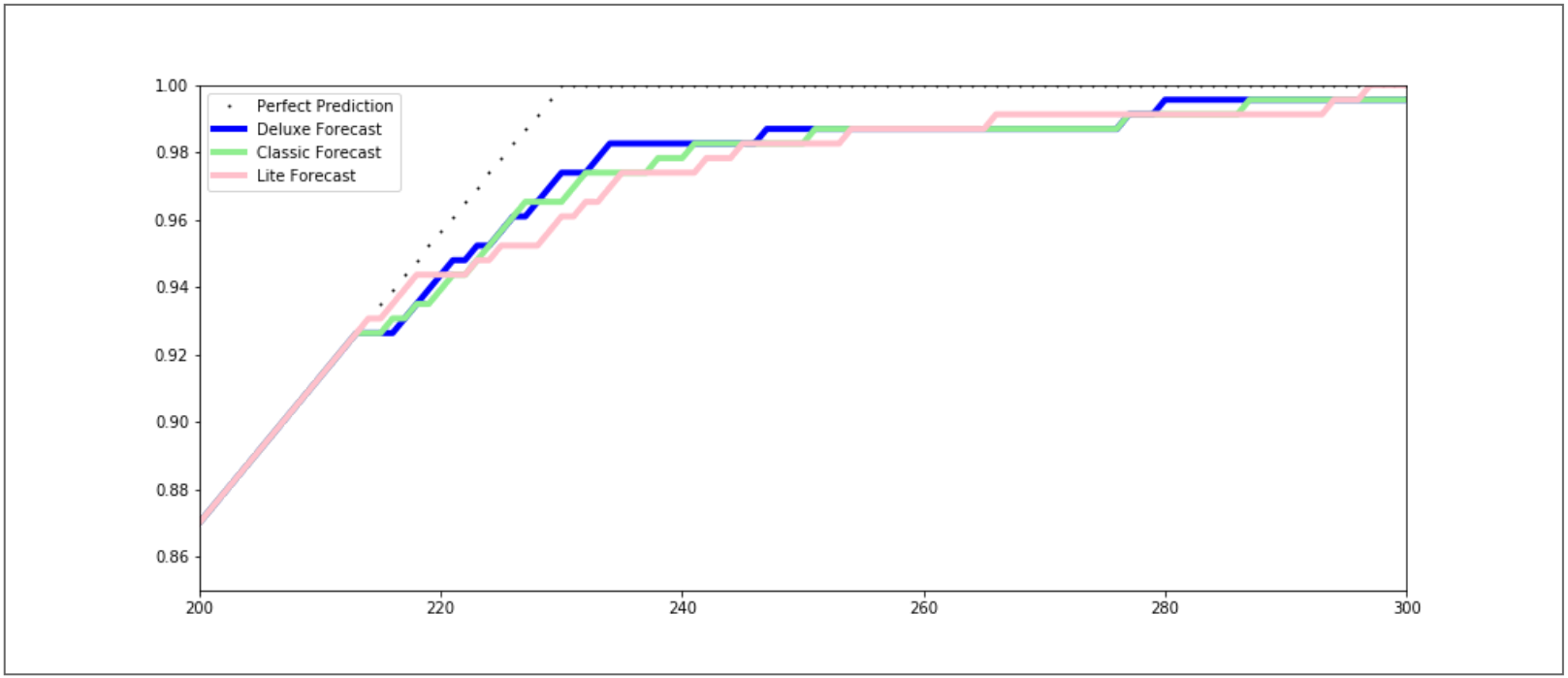
538 Lift


Perfect Prediction
November Forecast
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
200
220
240
260
280
300

538 Lift


Perfect Prediction
November Forecast
August Forecast
September Forecast
October Forecast
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
200
220
240
260
280
300

538 Lift


Perfect Prediction
Deluxe Forecast
Classic Forecast
Lite Forecast
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
200
220
240
260
280
300

ROC Curves

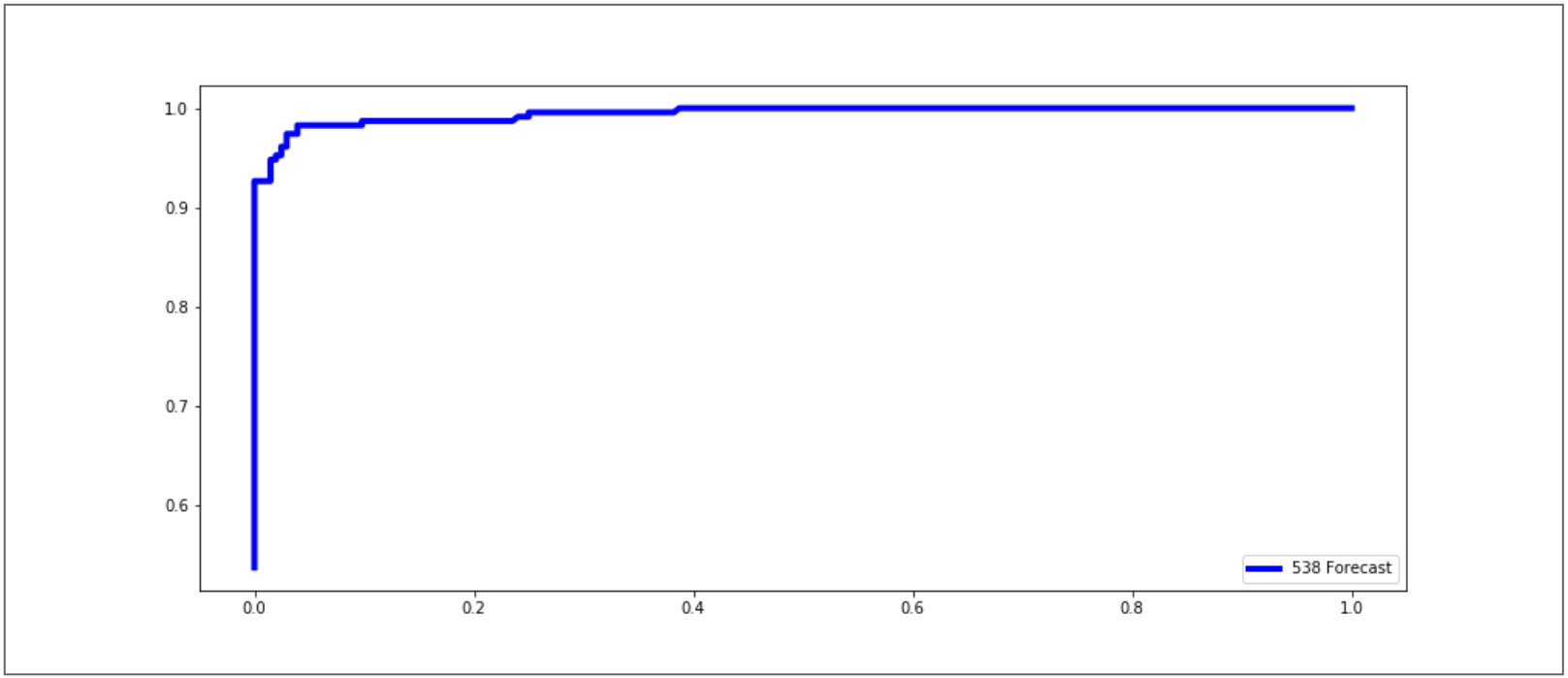
ROC = Receiver Operating Characteristic

Plot the true positive rate against the false positive rate at every possible threshold from highest to lowest.

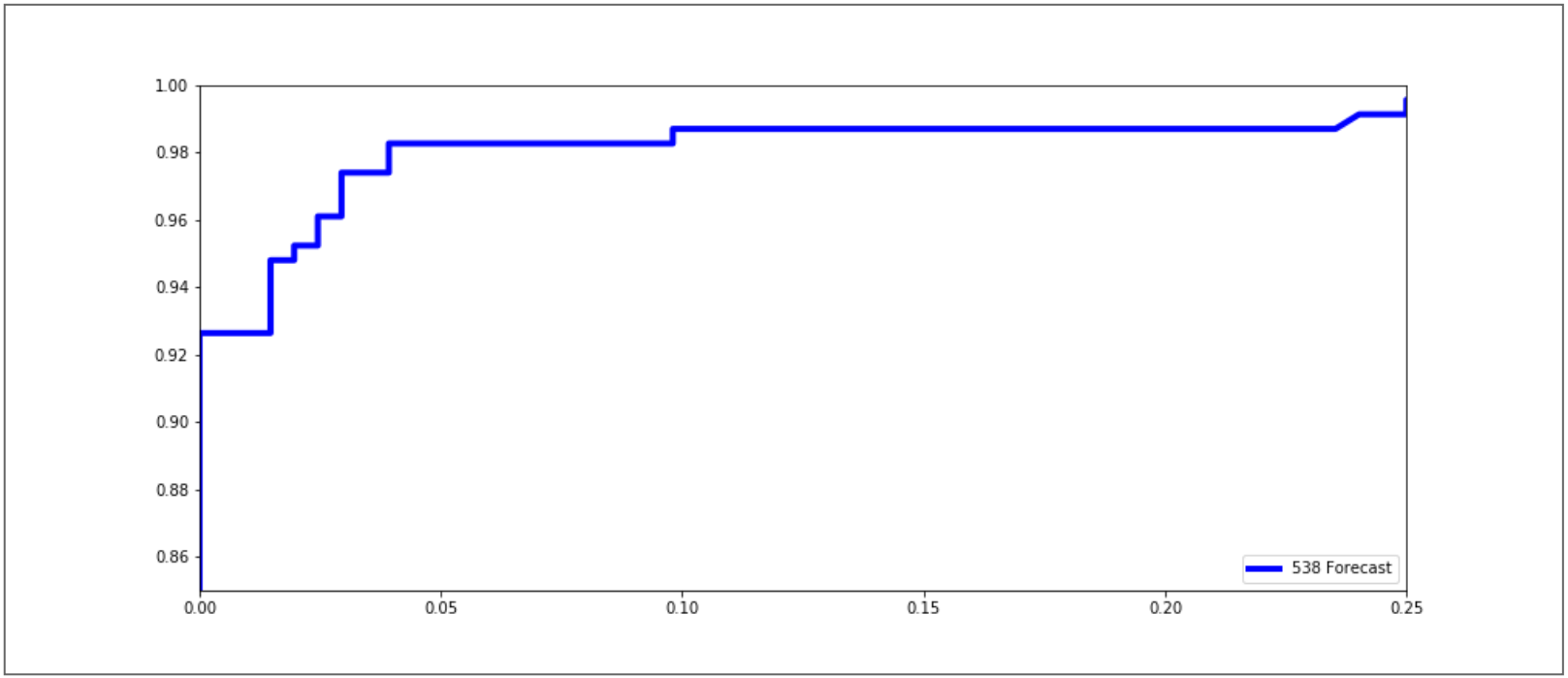
Demonstrates model's ability to outperform other (random) choices across decision thresholds while weighing false positives against false negatives.

For more details.

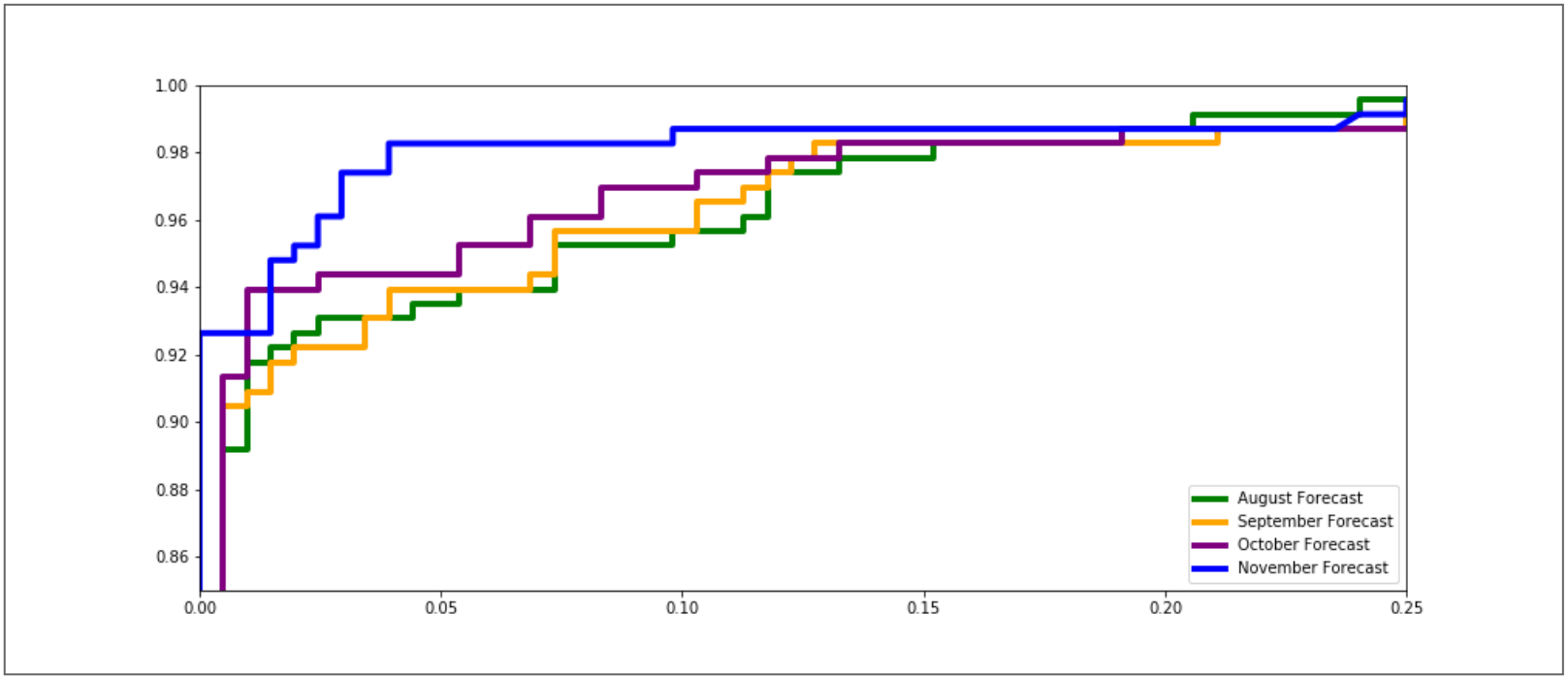
538 ROC



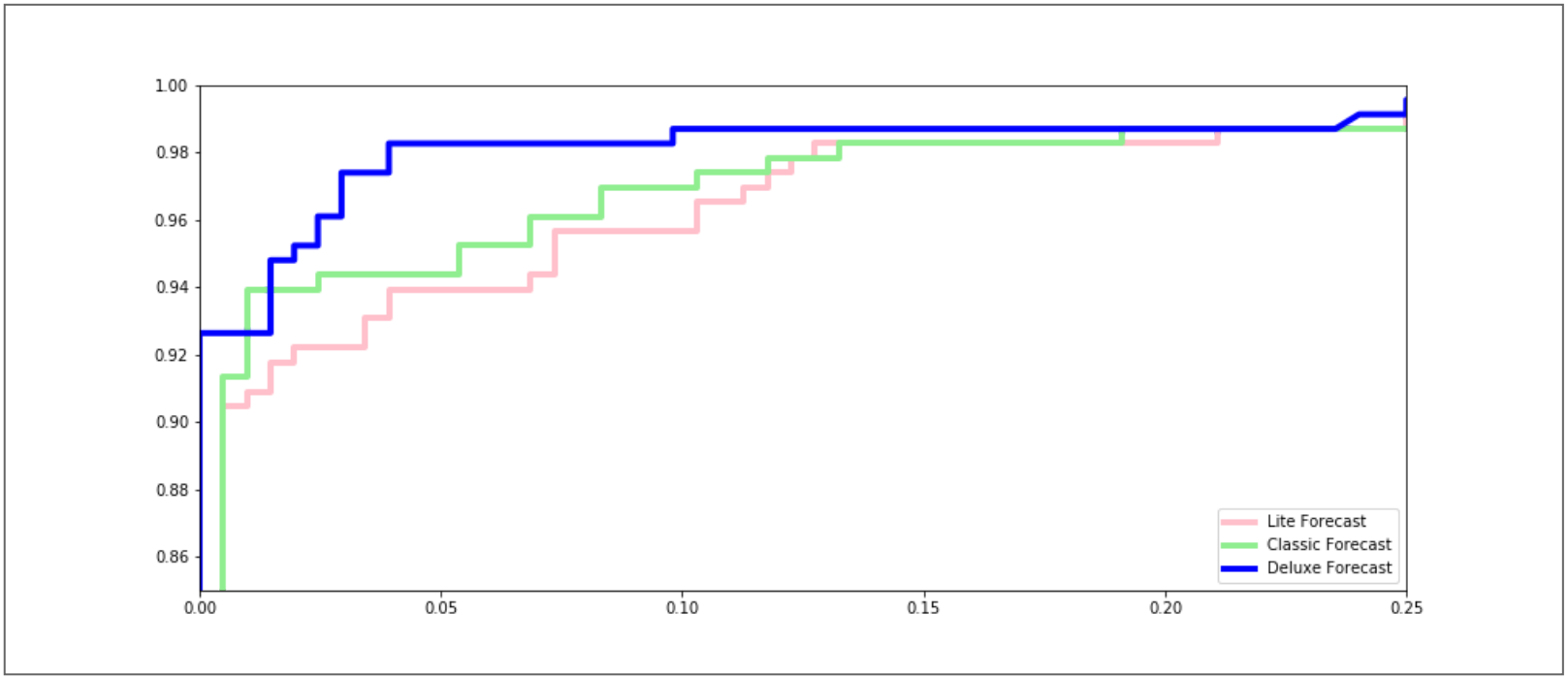
538 ROC


1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.00
0.05
0.10
0.15
0.20
0.25
538 Forecast

538 ROC


1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.00
0.05
0.10
0.15
0.20
0.25
August Forecast
September Forecast
October Forecast
November Forecast

538 ROC


1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.00
0.05
0.10
0.15
0.20
0.25
Lite Forecast
Classic Forecast
Deluxe Forecast

Can we reduce this to a single number?

AUC

AUC = Area Under the Curve

Total volume of area under the ROC curve.

Sci-kit Learn can calculate this for you.

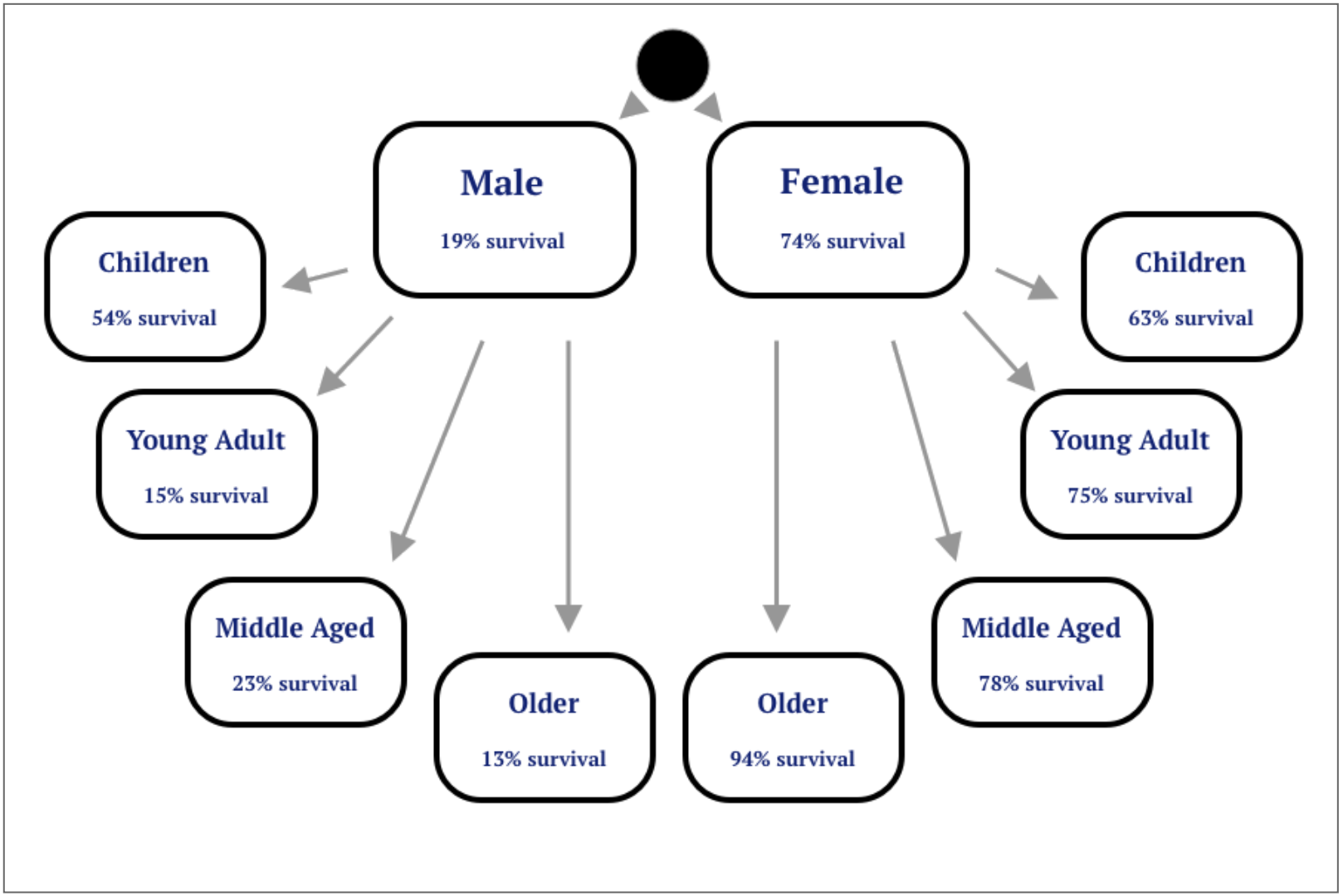
Part II

Ensemble Models

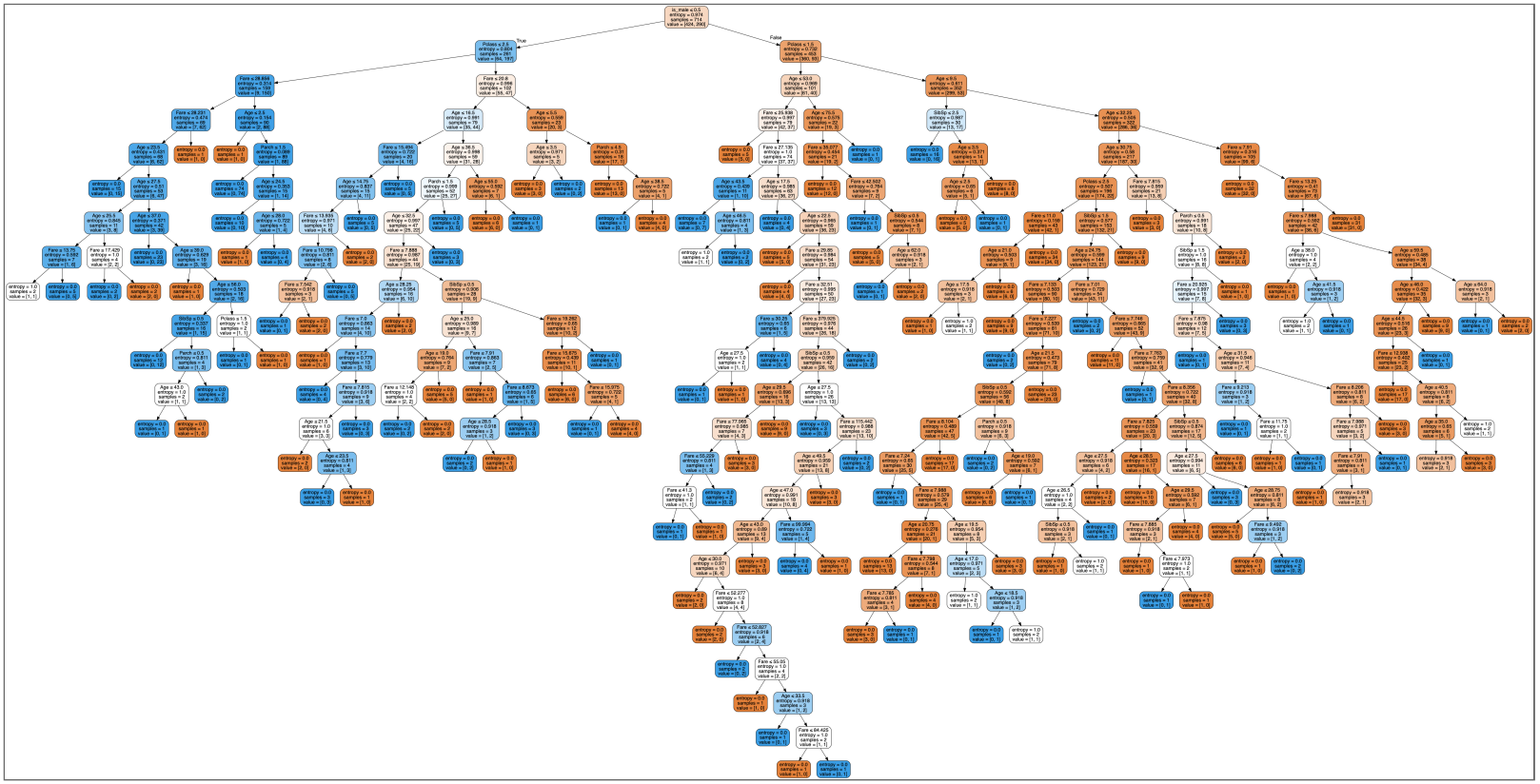
Classification Models

1. Logistic Regression
2. Decision Trees
3. Random Forest
4. Gradient Boosting Machine
5. Naive Bayes

Review: Survival Rates Among Subgroups

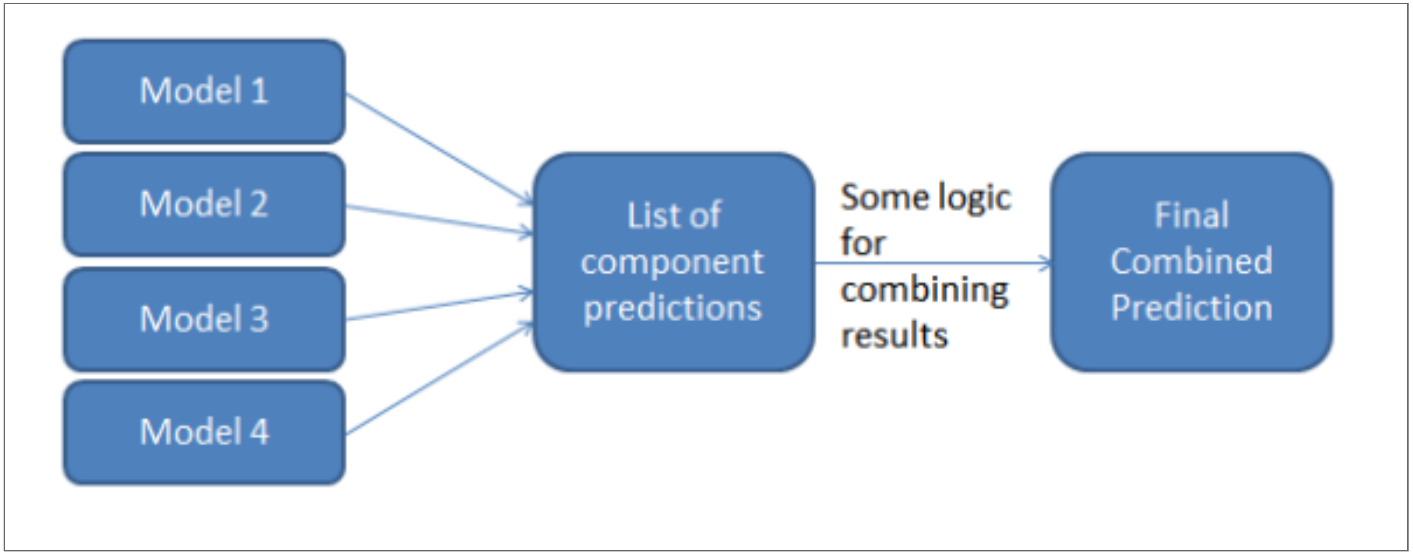


Review: Survival Rates Among Subgroups



Ensemble Models

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.



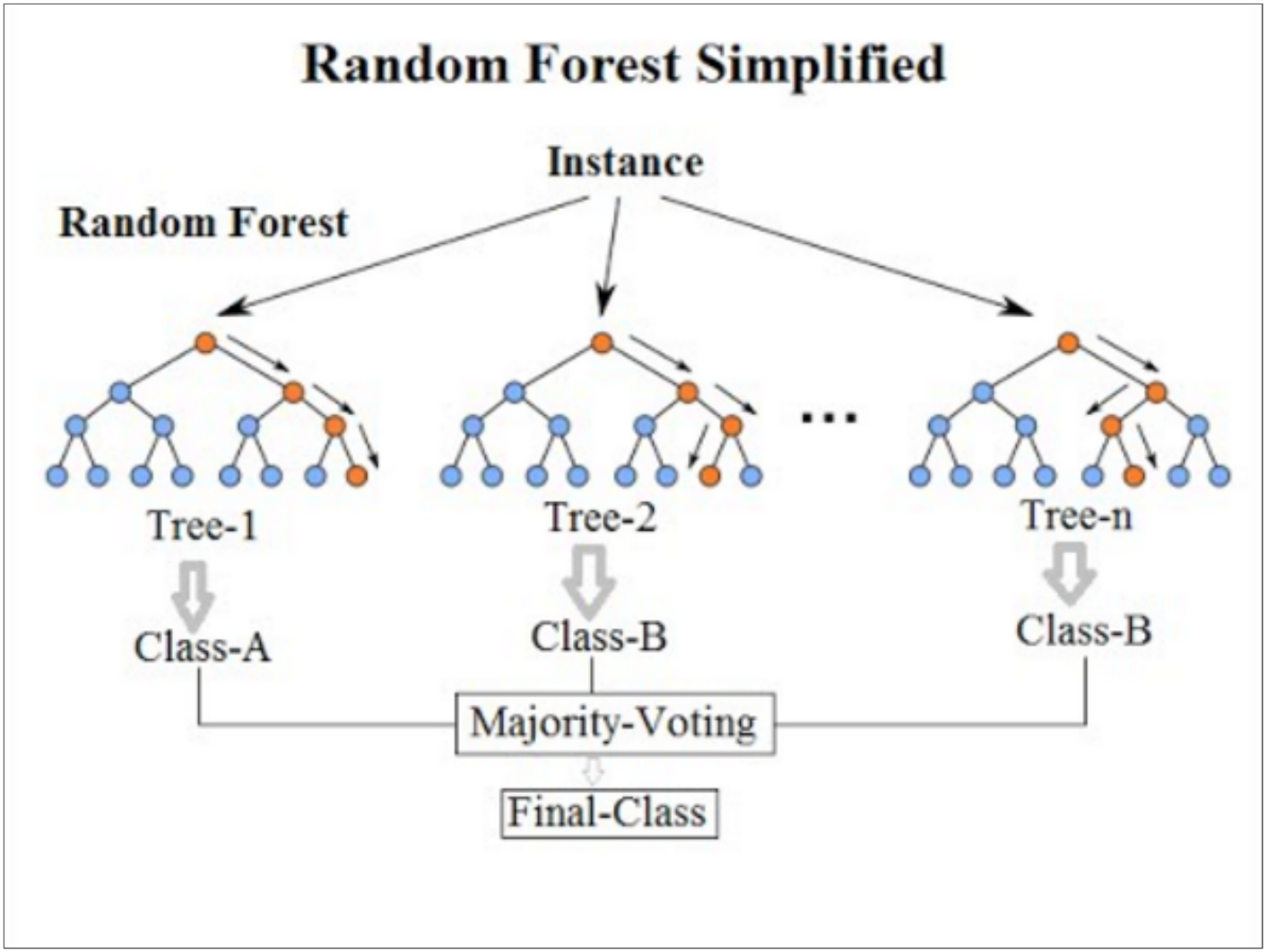
Random Forests

Random Forests construct a multitude of decision trees at training time and outputting the class that is the mode of the predicted classes among each tree.

For more details.

And a nice overview of everything.

Random Forests



Gradient Boosting

Gradient Boosting produces a prediction model in the form of an ensemble of weak prediction models and then generalizes them by allowing optimization of an arbitrary loss function.

For more details.

Gradient Boosting

Demo

Let's try.